Report

Overview:

The purpose of this analysis is to build and optimize a deep learning model to predict whether an organization will successfully use the funding it receives. Using historical data on applications, we preprocess and feed the dataset into a neural network, adjusting its architecture to maximize predictive accuracy.

Results:

Data Preprocessing

- Target Variable:
  - IS_SUCCESSFUL → The binary classification target (1 = successful, 0 = unsuccessful).
- Feature Variables:
  - All other columns except the removed ones (see below).
  - Features include APPLICATION_TYPE, AFFILIATION, CLASSIFICATION, USE_CASE, ORGANIZATION, INCOME_AMT, SPECIAL_CONSIDERATIONS, ASK_AMT, etc.
- Removed Variables:
  - EIN → Identification number (not predictive).
  - NAME → Organization name (not predictive).
  - These were not relevant to the target prediction and could introduce unnecessary noise.

Preprocessing Steps:

Encoded categorical variables using one-hot encoding
Standardized numerical features using StandardScaler
Grouped rare categorical values into "Other" to improve generalization

---

Compiling, Training, and Evaluating the Model

Neural Network Architecture:

- Input Layer: Features from preprocessed dataset.
- Hidden Layers:
  - 128 neurons (LeakyReLU)
  - 64 neurons (LeakyReLU)

    - 32 neurons (LeakyReLU)
    - 16 neurons (LeakyReLU)
- Output Layer:
    - 1 neuron (Sigmoid activation for binary classification).

Activation Functions Used:

- LeakyReLU (to prevent dead neurons and improve gradient flow).
- Sigmoid (for binary classification).

To increase model accuracy, we added activation different functions, added more layers, and also added more epochs.

Recommendation for improvement in model prediction:

Despite multiple attempts of improving our model, the accuracy fell just short, landing at 72.43%. Since the neural network was unable to reach 75% accuracy, a different machine learning approach may be more effective. Random Forest is well-suited for this classification as it effectively handles structured, tabular data with a mix of categorical and numerical features. Unlike deep learning, which requires extensive tuning and large datasets, RF performs well with limited data by reducing overfitting. It can also manage imbalanced classes and provide feature importance insights, helping to identify the most influential factors in determining funding success.Hence, Random Forest is a more practical choice for improving model accuracy in this scenario.